# Supplementary Material: A Prior-Less Method for Multi-Face Tracking in Unconstrained Videos

Chung-Ching Lin
IBM Research AI
cclin@us.ibm.com

Ying Hung
Rutgers University
yhung@stat.rutgers.edu

## 1. Overview

This supplementary material presents mathematical details of Gaussian Process and extends the experimental section presented in the main manuscript.

1. **Algorithmic details**. We describe more algorithmic details and experimnts for Section 3.3 Refinement Based on Gaussian Process (GP) of the main paper.
2. **Quantitative evaluation**. We provide the statistics of both datasets, the definitions of evaluation metrics, and quantitative multi-face tracking results on each video in Section 3.
3. **Qualitative evaluation**. We present sample qualitative tracking results in Section 4.

## 2. Algorithmic Details

### 2.1. Asymptotic consistency of GP

In Section 3.3 of the main manuscript, Gaussian Process (GP) models are used to detect outliers. GP modeling is an efficient parametric approach commonly used in spatial statistics and machine learning. Because the information in each image is spatially correlated, a reasonable assumption is that the data in each image is a realization of a GP with unknown parameters:

$$y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{16}$$

According to Mardia and Marshall (1984), the MLE obtained by Equations (13)-(15) are consistent estimators of the underlying true parameters. Define $\widehat{\gamma} = (\widehat{\boldsymbol{\theta}}', \widehat{\sigma}^2)'$ and $\widehat{\phi} = (\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\gamma}}')'$, the follow theoretical properties can be obtained by simple modification of Mardia and Marshall (1984).

**Remark 1:** Under some regularity conditions given by Mardia and Marshall (1984), $\widehat{\phi}$ is a consistent estimator of the true $\phi$ and it is asymptotically normal distributed. That is

$$\widehat{\phi} \sim N(\phi, B_n^{-1}), \tag{17}$$

where

$$B_n = diag(B_{\boldsymbol{\beta}}, B_{\boldsymbol{\gamma}}), B_{\boldsymbol{\beta}} = X' \frac{\Sigma^{-1}(\boldsymbol{\theta})}{\sigma^2} X, \tag{18}$$

and the $(i,j)$th element of $B_{\boldsymbol{\gamma}}$ is $\frac{1}{2} t_{ij}$ with

$$t_{ij} = trace(\sigma^2 \Sigma(\boldsymbol{\theta}) V^i \sigma^2 \Sigma(\boldsymbol{\theta}) V^j), \tag{19}$$

where

$$V^i = -\frac{\Sigma^{-1}(\boldsymbol{\theta})}{\sigma^2} V_i \frac{\Sigma^{-1}(\boldsymbol{\theta})}{\sigma^2},\ V_1 = \sigma^2 \frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \theta_1},\ V_2 = \sigma^2 \frac{\partial \Sigma(\boldsymbol{\theta})}{\partial \theta_2},\ V_3 = \Sigma(\boldsymbol{\theta}),\ i,j = 1, ..., 3. \tag{20}$$

Based on the theoretical support from Remark 1, the fitted GP successfully captures the information of each image with a much lower dimension, i.e., 18 parameters in total. Therefore, outlier detection can be efficiently and effectively performed based on the fitted GP parameters.

### 2.2. Examples of Outlier Detection

Figure 4(a) shows an initial cluster in foot chase video. Each small image is a sample from a face tracklet. We can see the variations of face appearances caused by poor lighting and severe camera motion.

Figure 4(b) shows the outlier detection results. From Figure 4(b), we can see that all outliers of the cluster in Figure 4(a) are detected and only one detection error exists. In general, clustering errors are caused by small, profile, or occluded faces that cannot be distinguished by deep face features. GP model is able to compensate the insufficiency of the CNN-based initialized linking framework and capture the false positive tracklet associations.

(a)


(b)

Figure 4: Outlier detection example in Foot Chase video: (a) an initial cluster with noise (b) detected outliers

## 3. Quantitative Evaluation

In our experiments, we performed the methods in [7] using the source code provided by the authors. (https://github.com/shunzhang876/AdaptiveFeatureLearning).

### 3.1. Evaluation Metrics

We report tracking results based on the most widely accepted evaluation metrics, the CLEAR MOT [6], including Recall, Precision, F1, FAF, MT, IDS, Frag, MOTA, and MOTP. The definitions are listed in Table 5. The up arrows indicate higher scores are better and vice versa.

### 3.2. Dataset Statistics

The characteristics of the music video dataset provided by [7] and body-worn camera datasets provided by our paper are shown in Table 6. The body-worn camera dataset brings different challenges from music video dataset. The camera movement is unstable since the camera is mounted on the human bodies. Take the Foot Chase video as an example, the video is about polices chasing and catching a suspect, thus the video has severe camera movements. The video can be found in the link: https://www.youtube.com/watch?v=StBOrFouFmE.

Table 5: Evaluation metrics for multi-face tracking. The up arrows indicate higher scores are better and vice versa.

| Name | Definition |
|---|---|
| Recall↑ | (Frame-based) correctly matched objects / total ground truth objects |
| Precision↑ | (Frame-based) correctly matched objects / total output objects |
| F1↑ | The harmonic mean of precision and recall. $F1 = 2(Precision * Recall)/Precision + Recall)$ |
| FAF↓ | (Frame-based) No. of false alarms per frame |
| GT | No. of ground truth trajectories |
| MT↑ | Mostly tracked: Percentage of GT trajectories which are covered by tracker output for more than 80% in length |
| PT↓ | Partially tracked: Percentage of GT trajectories which are covered by tracker output for less than 80% in length and more than 20% |
| Frag↓ | Fragments: The total of No. of times that a ground truth trajectory is interrupted in tracking result |
| IDS↓ | ID switches: The total of No. of times that a tracked trajectory changes its matched GT identity |
| MOTA↑ | The Multiple Object Tracking Accuracy takes into account false positives, missed targets and identity switches |
| MOTP↑ | The Multiple Object Tracking Precision is simply the average distance between true and estimated targets |

Table 6: Statistics of datasets.

Music Video Dataset

| Video | Duration(sec) | Frames | Main casts | Number of faces | Resolution |
|---|---|---|---|---|---|
| T-ara | 152 | 4,547 | 6 | 12,595 | 1280x720 |
| Pussycat Dolls | 198 | 5,937 | 6 | 17,515 | 1280x720 |
| Bruno Mars | 270 | 6,483 | 11 | 14,837 | 1280x720 |
| Hello Bubble | 220 | 5,275 | 4 | 4,731 | 1280x720 |
| Darling | 157 | 3,769 | 8 | 11,522 | 1280x720 |
| Apink | 197 | 4,729 | 6 | 6,294 | 1280x720 |
| Westlife | 229 | 5,736 | 4 | 27,306 | 1280x720 |
| Girls Aloud | 221 | 5,531 | 5 | 22,798 | 854x480 |

Body-worn Camera Dataset

| Video | Duration(sec) | Frames | Main casts | Number of faces | Resolution |
|---|---|---|---|---|---|
| Foot Chase | 762 | 12,076 | 5 | 5207 | 640x480 |
| TS1 | 128 | 1,807 | 2 | 631 | 640x480 |
| TS3 | 35 | 1,027 | 3 | 200 | 640x480 |
| DVHD2 | 266 | 7,981 | 3 | 1137 | 1280x720 |

### 3.3. Quantitative Multi-face Tracking Results

We report the face tracking results on each music video in Table 7 and 8. Our method is compared with ADMM[1], IHTLS[3] and variants in [7]. Table 9 presents the face tracking results of our method and 4 variants in [7] on body-worn camera videos.

## 4. Qualitative Evaluation

Figures 5-16 show sample tracking results generated by our method in the Music Video and Body-worn Camera datasets.

Table 7: Quantitative comparison with the state-of-the-art methods on music video dataset. The best results are highlighted with the bold.

T-ARA

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 58.0 | 68.3 | 62.8 | 0.86 | 6 | **0** | **6** | 251 | 641 | 29.4 | 63.8 |
| IHTLS[3] | 58.0 | 73.2 | 64.7 | 0.68 | 6 | **0** | **6** | 218 | 632 | 35.3 | 63.8 |
| Pre-Trained[7] | 60.9 | 95.9 | 74.5 | 0.10 | 6 | **0** | **6** | 143 | 232 | 57.3 | 72.4 |
| mTLD[7] | 62.1 | 93.5 | 74.6 | 0.14 | 6 | **0** | **6** | 251 | 241 | 56.0 | 72.6 |
| Siamese[7] | 62.1 | 95.5 | 75.3 | 0.09 | 6 | **0** | **6** | 106 | 213 | 58.4 | 72.5 |
| Triplet[7] | **63.5** | 94.2 | **75.9** | 0.12 | 6 | **0** | **6** | 94 | 233 | 59.0 | 72.5 |
| SymTriplet[7] | 62.8 | 95.4 | 75.7 | 0.10 | 6 | **0** | **6** | **75** | 235 | **59.2** | 72.4 |
| Ours | 59.8 | **96.9** | 74.0 | **0.07** | 6 | **0** | **6** | 95 | **190** | 57.5 | **86.7** |

PUSSYCAT DOLLS

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 89.3 | 74.2 | 81.0 | 0.58 | 6 | **4** | **2** | 287 | 412 | 63.2 | 63.5 |
| IHTLS[3] | 89.5 | 78.6 | 83.7 | 0.42 | 6 | **4** | **2** | 248 | 413 | 70.3 | 63.5 |
| Pre-Trained[7] | 76.4 | 88.0 | 81.8 | 0.3 | 6 | 2 | 4 | 128 | 405 | 65.1 | 64.9 |
| mTLD[7] | 79.7 | **89.5** | 84.3 | **0.22** | 6 | 2 | 4 | 296 | 444 | 68.3 | 64.9 |
| Siamese[7] | 81.2 | 88.9 | **84.9** | 0.24 | 6 | 2 | 4 | 107 | 430 | **70.3** | 64.9 |
| Triplet[7] | 81.4 | 88.3 | 84.7 | 0.26 | 6 | 2 | 4 | 99 | 435 | 69.9 | 64.9 |
| SymTriplet[7] | **81.6** | 88.2 | 84.8 | 0.26 | 6 | 2 | 4 | 82 | 439 | 70.2 | 64.9 |
| Ours | 78.8 | 85.1 | 81.8 | 0.42 | 6 | 2 | 4 | **66** | **194** | 60.7 | **75.4** |

BRUNO MARS

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 68.9 | 76.0 | 72.3 | 0.40 | 11 | 3 | 8 | 428 | 503 | 50.6 | 85.7 |
| IHTLS[3] | 68.5 | 83.5 | 75.2 | 0.35 | 11 | 3 | 8 | 375 | 491 | 52.7 | 85.8 |
| Pre-Trained[7] | 53.7 | 92.3 | 67.9 | 0.10 | 11 | 0 | 9 | 151 | 453 | 48.3 | 88.0 |
| mTLD[7] | 58.0 | **94.0** | 71.7 | **0.10** | 11 | 2 | 9 | 278 | 551 | 52.6 | **87.9** |
| Siamese[7] | 62.3 | 92.8 | 74.6 | 0.12 | 11 | 2 | 8 | 126 | 540 | 56.7 | 87.8 |
| Triplet[7] | 62.4 | 92.6 | 74.6 | 0.13 | 11 | 2 | 9 | 126 | 543 | 56.6 | 87.8 |
| SymTriplet[7] | 62.9 | 91.9 | 74.7 | 0.14 | 11 | 2 | 9 | **105** | 551 | 56.8 | 87.8 |
| Ours | **84.7** | 85.7 | **85.1** | 0.48 | 11 | **8** | **3** | 220 | **501** | **65.8** | 82.0 |

HELLO BUBBLE

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 66.1 | 80.2 | 72.5 | 0.23 | 4 | **0** | **4** | 115 | 191 | 47.6 | 69.9 |
| IHTLS[3] | 65.9 | 84.8 | 74.2 | 0.16 | 4 | **0** | **4** | 109 | 190 | 52.0 | 69.9 |
| Pre-Trained[7] | 47.1 | 83.8 | 60.3 | 0.10 | 4 | **0** | **4** | 71 | 187 | 36.6 | 68.5 |
| mTLD[7] | 67.4 | 84.8 | 75.1 | 0.17 | 4 | **0** | **4** | 139 | 255 | 52.6 | 70.5 |
| Siamese[7] | 67.6 | 88.0 | 76.5 | **0.13** | 4 | **0** | **4** | 105 | 249 | 56.3 | 70.6 |
| Triplet[7] | 68.6 | 86.4 | 76.5 | 0.15 | 4 | **0** | **4** | 82 | 256 | 56.2 | 70.5 |
| SymTriplet[7] | 68.6 | 86.5 | 76.5 | 0.15 | 4 | **0** | **4** | 69 | 256 | 56.5 | 70.5 |
| Ours | **73.3** | **89.2** | **80.5** | 0.17 | 4 | **0** | **4** | **51** | **148** | **60.9** | **84.8** |

Table 8: Quantitative comparison with the state-of-the-art methods on music video dataset. The best results are highlighted with the bold.

DARLING

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 88.3 | 74.0 | 80.6 | 0.62 | 8 | **7** | **1** | 412 | 342 | 53.0 | 88.4 |
| IHTLS[3] | 88.5 | 80.2 | 84.2 | 0.44 | 8 | **7** | **1** | 381 | 338 | 62.7 | 88.4 |
| Pre-Trained[7] | 53.1 | 85.2 | 65.4 | 0.20 | 8 | 2 | 6 | 115 | 233 | 42.7 | 88.5 |
| mTLD[7] | 79.9 | 82.3 | 81.1 | 0.35 | 8 | 4 | 4 | 278 | 461 | 59.8 | **89.3** |
| Siamese[7] | 85.2 | 86.3 | 85.7 | 0.27 | 8 | **7** | **1** | 214 | 310 | 69.5 | 88.9 |
| Triplet[7] | 85.9 | 85.3 | 85.6 | 0.30 | 8 | **7** | **1** | 187 | 317 | 69.2 | 88.9 |
| SymTriplet[7] | 86.7 | 85.7 | 86.2 | 0.29 | 8 | **7** | **1** | 169 | 323 | 70.5 | 88.9 |
| Ours | **89.6** | **92.5** | **91.0** | **0.14** | 8 | **7** | **1** | **98** | **211** | **82.3** | 88.9 |

APINK

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 81.2 | 92.8 | 86.6 | 0.09 | 6 | 4 | 2 | 179 | 158 | 72.4 | 76.1 |
| IHTLS[3] | 81.2 | 95.4 | 87.7 | 0.05 | 6 | 4 | 2 | 173 | 157 | 74.9 | 76.1 |
| Pre-Trained[7] | 56.4 | 98.3 | 71.7 | 0.01 | 6 | 0 | 6 | 100 | 170 | 54.0 | 75.5 |
| mTLD[7] | 81.5 | 98.0 | 89.0 | 0.02 | 6 | 3 | 3 | 173 | 240 | 77.4 | 76.3 |
| Siamese[7] | 81.6 | **98.9** | 89.4 | **0.01** | 6 | 3 | 3 | 124 | 238 | 79.0 | 76.3 |
| Triplet[7] | 82.1 | 98.5 | 89.6 | 0.02 | 6 | 4 | 2 | 140 | 244 | 78.9 | 76.3 |
| SymTriplet[7] | 82.4 | 98.3 | 89.7 | 0.02 | 6 | 4 | 2 | 78 | 246 | 80.0 | 76.3 |
| Ours | **90.3** | 93.4 | **91.8** | 0.10 | 6 | **6** | **0** | **36** | **131** | **82.7** | **94.3** |

WESTLIFE

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 89.1 | 36.0 | 51.3 | 0.60 | 4 | **4** | **0** | 223 | 184 | 62.4 | 87.5 |
| IHTLS[3] | 89.4 | 39.9 | 55.2 | 0.65 | 4 | **4** | **0** | 113 | 177 | 60.9 | 87.5 |
| Pre-Trained[7] | 77.8 | 79.5 | 78.6 | 0.40 | 4 | 1 | 3 | 85 | 128 | 57.0 | 88.2 |
| mTLD[7] | 86.0 | 76.5 | 81.0 | 0.52 | 4 | 3 | 1 | 177 | 169 | 58.1 | 88.1 |
| Siamese[7] | 86.8 | 79.7 | 83.1 | 0.44 | 4 | 3 | 1 | 74 | 142 | 64.1 | 88.0 |
| Triplet[7] | 86.8 | 80.1 | 83.3 | 0.43 | 4 | 3 | 1 | 89 | 140 | 64.5 | 88.0 |
| SymTriplet[7] | 85.6 | 83.9 | 84.7 | **0.33** | 4 | 3 | 1 | 57 | 136 | 68.6 | 88.1 |
| Ours | **91.2** | **85.7** | **88.4** | 0.35 | 4 | **4** | **0** | **16** | **109** | **73.2** | **89.1** |

GIRLSALOUD

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ADMM[1] | 70.0 | 50.3 | 58.5 | 0.61 | 5 | 1 | 4 | 487 | 528 | 46.6 | 87.1 |
| IHTLS[3] | 69.8 | 60.2 | 64.7 | 0.46 | 5 | 1 | 4 | 396 | 482 | 51.8 | 87.2 |
| Pre-Trained[7] | 49.3 | 89.6 | 63.6 | 0.20 | 5 | 0 | 5 | 138 | 332 | 42.7 | 87.7 |
| mTLD[7] | 54.3 | 90.5 | 67.9 | 0.17 | 5 | 0 | 5 | 322 | 425 | 46.7 | **88.2** |
| Siamese[7] | 58.1 | 90.8 | 70.9 | 0.17 | 5 | 1 | 4 | 112 | 376 | 51.6 | 87.8 |
| Triplet[7] | 57.2 | 92.0 | 70.5 | **0.15** | 5 | 1 | 4 | 80 | 367 | 51.7 | 87.8 |
| SymTriplet[7] | 58.2 | 90.3 | 70.8 | 0.19 | 5 | 1 | 4 | 64 | 377 | 51.6 | 87.8 |
| Ours | **86.0** | **93.1** | **89.4** | 0.42 | 5 | **5** | **0** | **42** | **161** | **71.5** | 87.3 |

Table 9: Quantitative comparisons with the state-of-the-art method [7] on body-worn camera dataset.

Foot Chase

| Method | Rec.↑ | Prec.↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mTLD[7] | 71.5 | 84.5 | 77.5 | 0.09 | 5 | 2 | 3 | 51 | 272 | 50.9 | 93.2 |
| Pre-Trained[7] | 71.5 | 84.5 | 77.5 | 0.09 | 5 | 2 | 3 | 43 | 271 | 51.0 | 93.2 |
| Siamese[7] | 71.4 | 84.5 | 77.4 | 0.09 | 5 | 2 | 3 | 38 | 275 | 51.1 | 93.2 |
| SymTriplet[7] | 71.4 | 84.9 | 77.6 | 0.09 | 5 | 2 | 3 | 32 | 271 | 51.6 | 93.2 |
| Ours | **76.6** | **95.7** | **85.1** | **0.01** | 5 | **5** | **0** | **30** | **155** | **73.2** | **94.3** |

TS1

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mTLD[7] | 87.3 | 43.2 | 57.8 | 0.34 | 2 | 1 | 1 | 7 | 18 | 67.0 | 95.9 |
| Pre-Trained[7] | 87.3 | 43.2 | 57.8 | 0.34 | 2 | 1 | 1 | 7 | 18 | 67.0 | 95.9 |
| Siamese[7] | 87.3 | 43.2 | 57.8 | 0.34 | 2 | 1 | 1 | 6 | 18 | 67.1 | 96.0 |
| SymTriplet[7] | **87.3** | 43.2 | 57.8 | 0.34 | 2 | **1** | **1** | 4 | 18 | 67.5 | **95.9** |
| Ours | 81.5 | **86.2** | **83.8** | **0.22** | 2 | **1** | **1** | **4** | **8** | **68.5** | 94.0 |

TS3

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mTLD[7] | 88.5 | 35.1 | 50.3 | 0.32 | 3 | 2 | 1 | 2 | 32 | 55.9 | 81.7 |
| Pre-Trained[7] | 88.0 | 34.6 | 49.7 | 0.34 | 3 | 2 | 1 | 2 | 34 | 52.0 | 81.8 |
| Siamese[7] | 88.5 | 36.0 | 51.2 | 0.29 | 3 | 2 | 1 | 2 | 28 | 58.4 | 81.6 |
| SymTriplet[7] | 88.5 | 35.0 | 50.2 | 0.34 | 3 | **2** | 1 | 2 | 33 | 53.1 | **81.8** |
| Ours | **95.0** | **80.9** | **87.4** | **0.30** | 3 | **3** | **0** | **1** | **9** | **64.5** | 75.3 |

DVHD2

| Method | Recall↑ | Precision↑ | F1↑ | FAF↓ | GT | MT↑ | PT↓ | IDS↓ | Frag↓ | MOTA↑ | MOTP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mTLD[7] | 82.4 | 82.9 | 82.6 | 0.21 | 3 | 2 | 1 | 10 | 78 | 52.5 | 95.6 |
| Pre-Trained[7] | 82.9 | 82.7 | 82.8 | 0.20 | 3 | 2 | 1 | 9 | 81 | 54.0 | 95.3 |
| Siamese[7] | 82.9 | 82.6 | 82.7 | 0.21 | 3 | 2 | 1 | 8 | 83 | 52.7 | 95.3 |
| SymTriplet[7] | 82.8 | 84.7 | 83.7 | **0.18** | 3 | **2** | **1** | 11 | 68 | **57.2** | **95.3** |
| Ours | **83.5** | **91.6** | **87.4** | 0.20 | 3 | **2** | **1** | **4** | **16** | 55.8 | 93.6 |


Figure 5: Sample tracking results of the proposed algorithm on T-ara music video. The ID number and color of face bounding box for each person are kept.

Figure 6: Sample tracking results of the proposed algorithm on Pussycat Dolls music video. The ID number and color of face bounding box for each person are kept.


Figure 7: Sample tracking results of the proposed algorithm on BrunoMars music video. The ID number and color of face bounding box for each person are kept.


Figure 8: Sample tracking results of the proposed algorithm on HelloBubble music video. The ID number and color of face bounding box for each person are kept.


Figure 9: Sample tracking results of the proposed algorithm on Darling music video. The ID number and color of face bounding box for each person are kept.


Figure 10: Sample tracking results of the proposed algorithm on Apink music video. The ID number and color of face bounding box for each person are kept.


Figure 11: Sample tracking results of the proposed algorithm on Westlife music video. The ID number and color of face bounding box for each person are kept.


Figure 12: Sample tracking results of the proposed algorithm on GirlsAloud music video. The ID number and color of face bounding box for each person are kept.

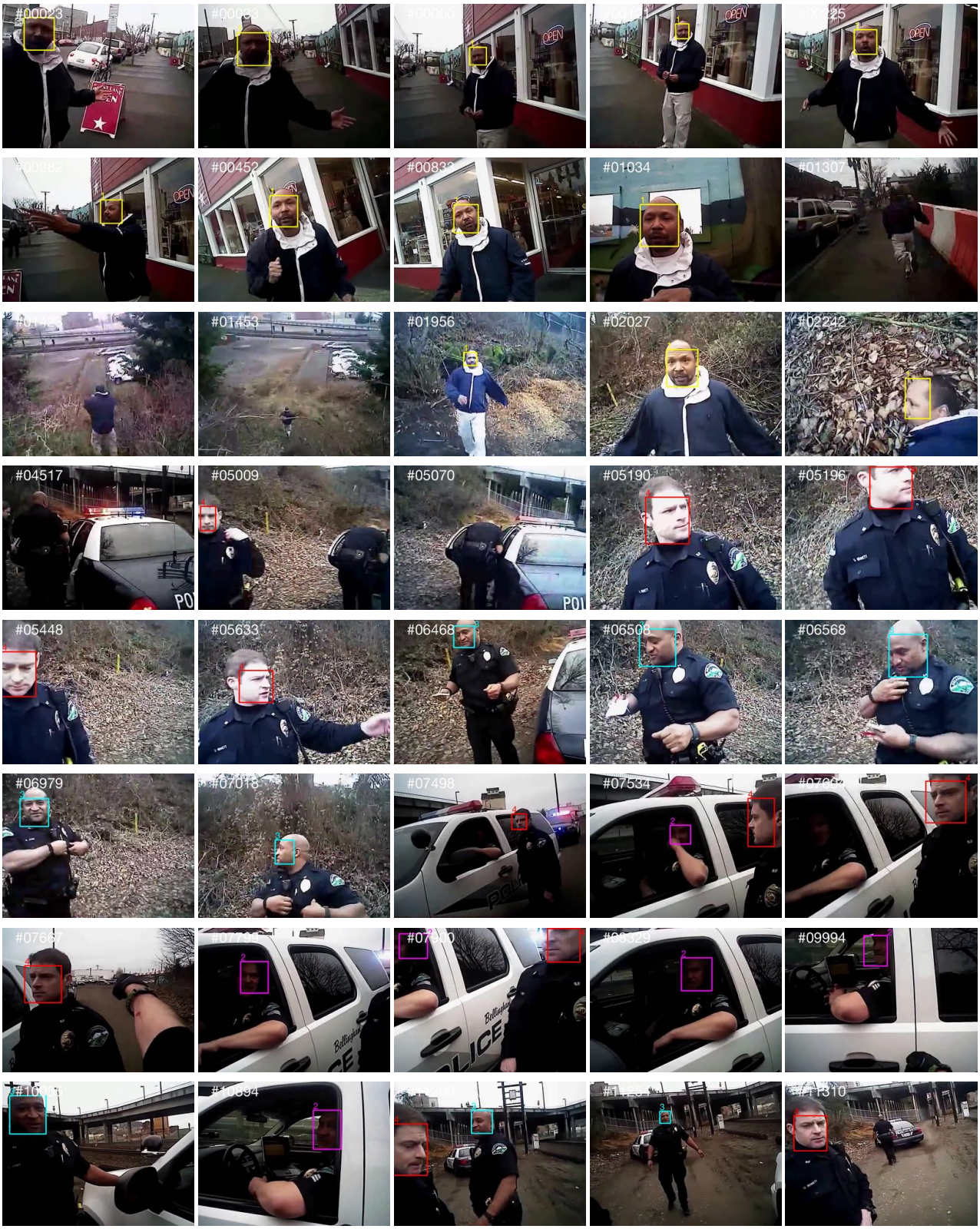


Figure 13: Sample tracking results of the proposed algorithm on Foot Chase video. Sample tracking results of the proposed algorithm


Figure 14: Sample tracking results of the proposed algorithm on TS1 video. Sample tracking results of the proposed algorithm

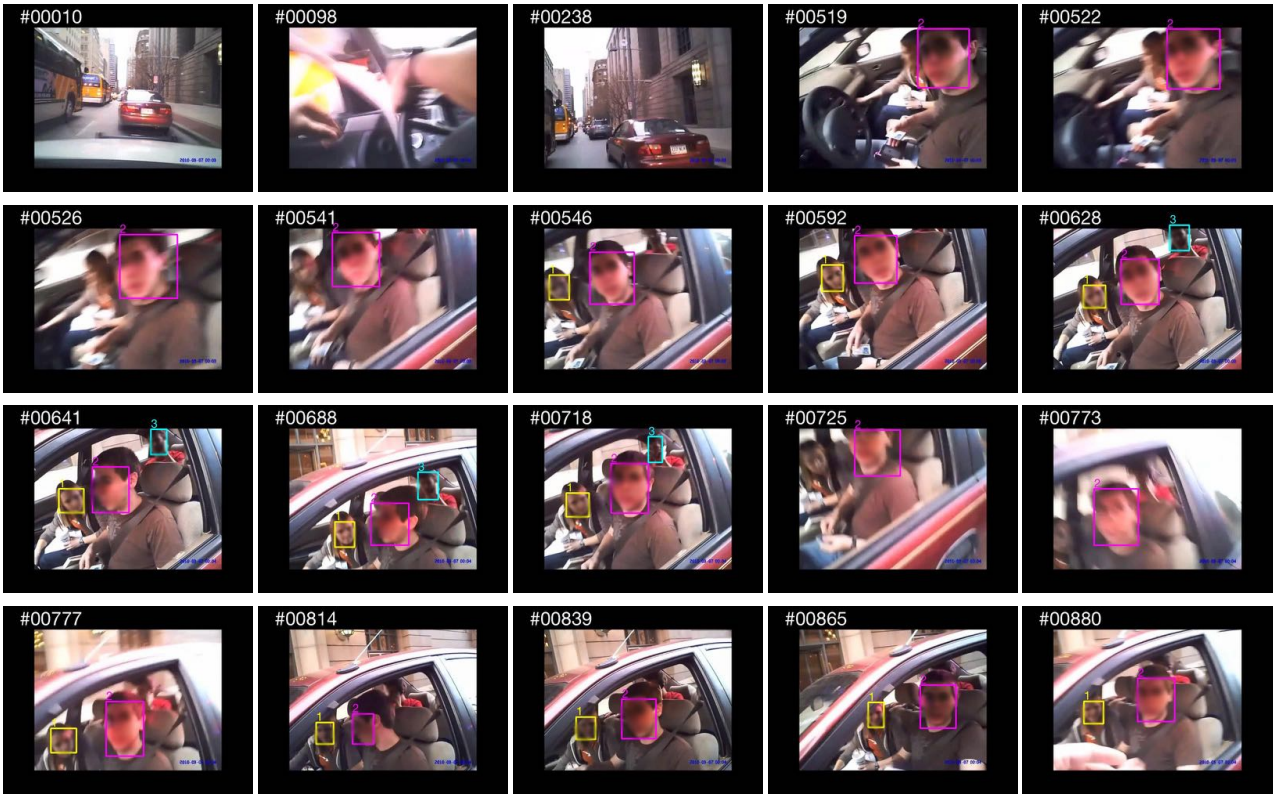


Figure 15: Sample tracking results of the proposed algorithm on Traffic Stop 3 video. Sample tracking results of the proposed algorithm


Figure 16: Sample tracking results of the proposed algorithm on Domestic Violence HD2 music video. Sample tracking results of the proposed algorithm

## References


[1] M. Ayazoglu, M. Sznaier, and O. I. Camps. Fast algorithms for structured robust principal component analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1704–1711. IEEE, 2012. 3, 4, 5

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[3] C. Dicle, O. I. Camps, and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2311, 2013. 3, 4, 5

[4] T. Hastie, R. I. Tibshirani, et al. The elements of statistical learning: data mining, inference, and prediction/by trevor hastie, robert tibshirani, jerome frieman. Technical report, 2009.

[5] G. James, D. Witten, and T. Hastie. An introduction to statistical learning: With applications in r., 2014.

[6] A. Milan, K. Schindler, and S. Roth. Challenges of ground truth evaluation of multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 735–742, 2013. 2

[7] S. Zhang, Y. Gong, J.-B. Huang, J. Lim, J. Wang, N. Ahuja, and M.-H. Yang. Tracking persons-of-interest via adaptive discriminative features. In *European Conference on Computer Vision*, pages 415–433. Springer, 2016. 2, 3, 4, 5, 6